Data Wrangling Report

Data Gathering:

There are 3 data sources in total.

1. Twitter_archive: This file is provided by Udacity. I used read_csv() from pandas library to load the data.
2. Image_predictions: This file **image_predictions.tsv** is downloaded programmatically by using request library to get the content from URL.
3. Tweet_data: Query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called **tweet-json.txt** file. For this step I copied the note from Udacity's instruction of getting Twitter API and used the provided file **tweet-json.txt** to continue my next steps.

Data Assessment:

I inspect the three data set based on their tidiness and quality measurements.

- ***Tidiness issues***
  1. Twitter_archive:
     - Columns 'doggo','floofer','pupper','puppo' does not need to be separated since they are about the same thing.
  2. All:
     - The three files/ data frames should be merged because they are part of the same observation unit (tweets).
- ***Quality issues***
  1. Twitter_archive:
     - There are rows that has value in column 'in_reply_to_status_id' and 'retweeted_status_id' because these records are retweets and replies to the original tweet. This shows duplicate content issue.
     - There are columns which have no value in our analysis purpose. Ex: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, expanded_urls
     - Column 'timestamp' is in object datatype.
     - The data is not consistent because there are 18 different denominators.
     - There are rows with rating denominator <= 0 - invalid data.
     - There are rows with rating numerator <= 0 - invalid data.
     - There are rows with rating denominator has extreme value - invalid data.
     - Some dogs' names are missing and incorrect. Ex: a, the…
     - The value in column 'source' are not easy to understand.
  2. Image_predictions:
     - There are columns that has no value in our analysis purpose.
     - There is no consistency in format of column p1, p2 and p3. Some values are capitalized, and others are lower case.
     - There are duplicated value of URL in Image Predictions data frame.

- These columns p1, p2 and p3 need to be more descriptive for reader to understand the content.
- There is only 1 URL link for rows which indicates there are more than one image.

Data Cleaning:

- ***Tidiness issues***
  1. Twitter_archive:
     - Columns 'doggo','floofer','pupper','puppo' are combined under one column named ‘TypeOfDog’.
  2. All:
     - Merge the three data sets into one big table by using 'tweet_id'.

- ***Quality issues***
  1. Twitter_archive:
     - Delete rows that has value in column 'in_reply_to_status_id' and 'retweeted_status_id' because it is retweet and replies to the original tweet.
     - Delete columns that has no value in our analysis purpose. Ex: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, expanded_urls
     - Column 'timestamp' is corrected to datetime data type.
     - Delete rows with rating denominator <= 0 - invalid data.
     - Delete rows with rating numerator <= 0 - invalid data.
     - Delete rows with rating numerator has extreme value - invalid data.
     - Delete rows with rating denominator has extreme value - invalid data.
     - Replace those incorrect dogs’ names to NaN.
     - Fix the value in column 'source' to be readable. Ex: Twitter for iPhone, Vine – Make a Scene, Twitter Web Client, TweetDeck.
  2. Image_predictions:
     - Delete columns that has no value in our analysis purpose. Ex: img_num
     - Change the value format of p1,p2 and p3 to be consistent. The first letter is capitalized and the rest is lower case.
     - Delete duplicated URL value from column ‘jpg_url’.
     - Change names in columns p1, p2 and p3 into ‘Algorithm 1 Prediction, Algorithm 2 Prediction, Algorithm 3 Prediction’.
     - There is only 1 URL link for rows which indicates there are more than one image. Delete column ‘img_num’.